scientific reports

Check for updates

OPEN

# Game-theoretic link relevance indexing on genome-wide expression dataset identifies putative salient genes with potential etiological and diapeutics role in colorectal cancer

Vishwa Jyoti Baruah[1]✉, Papori Neog Bora[2], Bhaswati Sarmah[3], Priyakshi Mahanta[4], Ankumon Sarmah[4], Stefano Moretti[5], Rajnish Kumar[6] & Surajit Borkotokey[2]✉

**Diapeutics gene markers in colorectal cancer (CRC) can help manage mortality caused by the disease. We applied a game-theoretic link relevance Index (LRI) scoring on the high-throughput whole-genome transcriptome dataset to identify salient genes in CRC and obtained 126 salient genes with LRI score greater than zero. The biomarkers database lacks preliminary information on the salient genes as biomarkers for all the available cancer cell types. The salient genes revealed eleven, one and six overrepresentations for major Biological Processes, Molecular Function, and Cellular components. However, no enrichment with respect to chromosome location was found for the salient genes. Significantly high enrichments were observed for several KEGG, Reactome and PPI terms. The survival analysis of top protein-coding salient genes exhibited superior prognostic characteristics for CRC. MIR143HG, AMOTL1, ACTG2 and other salient genes lack sufficient information regarding their etiological role in CRC. Further investigation in LRI methodology and salient genes to augment the existing knowledge base may create new milestones in CRC diapeutics.**

**Abbreviations**

| | |
|---|---|
| LRI | Link Relevance Index |
| CRC | Colorectal cancer |
| GO | Gene Ontology |
| BP | Biological Processes |
| MF | Molecular Function |
| CC | Cellular component |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| TCGA | The Cancer Genome Atlas |

Among the wide gamut of cancer types, colorectal cancer (CRC) marks its place as the third most recurrent type and seventh-most fatal disease/disorder globally[1,2]. Studies have shown that the prevalence of CRC increases with the person's age[3], though exceptional accounts of the disease occurrences in high proportion in the younger

[1]Centre for Biotechnology and Bioinformatics, Dibrugarh University, Dibrugarh 786004, Assam, India. [2]Department of Mathematics, Dibrugarh University, Dibrugarh 786004, Assam, India. [3]Department of Plant Breeding and Genetics, Assam Agricultural University, Jorhat 785013, Assam, India. [4]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh 786004, Assam, India. [5]CNRS, LAMSADE, Université Paris-Dauphine, PSL Research University, Paris, France. [6]Economics Group, Queen's Management School, Queen's University Belfast, Belfast BT9 5EE, UK. ✉email: vishwabaruah@gmail.com; sborkotokey@dibru.ac.in

population have also been reported[3,4]. Global CRC occurrence can be majorly attributed (95% cases) to factors other than a person's genetic predisposition and is a major contributor to cancer in developed and developing countries[1,3,5,6].

Though the philosophy of studying diagnostics and therapeutics in an intertwined manner is not new, a common umbrella term 'Diapeutics' under the larger domain of cancer covering both aspects was long overdue until recently[7]. The tumors confined to the colon region metastasize to nearby lymph nodes in the absence of an early diagnosis. Treatment of CRC varies from medication, chemotherapy in early detection to surgery, excision and specific targeted therapeutics in severe cases of metastases to the liver or lungs[6,8–11]. Hence, early diagnosis and screening hold the key to reducing the incidence and mortality caused by the disease. The colonoscopy procedure is the most widely used diagnosis and screening strategy; however, it has several shortcomings, including being cost-ineffective, invasive, non-reliable and precarious[8–11]. On the other hand, biomarkers address these limitations by presenting a more cost-effective, reliable, and non-invasive early detection and screening technique of CRC [8,11]. Salient genes as biomarkers can provide with the ability of prediction (e.g. BRAF, ALK, ROS1, HER2, PI3K and miR-31-3p), prognosis (e.g. CIMP, CDX2 and MYO5B) and diagnosis (e.g. KRAS, p53, EGFR, erbB2)[12,13].

Investigations on the tremendous amount of gene expression data generating specific patterns and gene co-expression networks and providing system-level functionality of genes have been effectively used to distinguish salient genes with the potential in diapeutics of a broad spectrum of complex human diseases[14]. The conventional data analysis methods on microarray take into account the down or upregulation of genes to find the salient genes. Only a few driver genes, when expressed aberrantly, are primarily responsible for the progression and advancement of the cancerous cell by bestowing the cell with a selective advantage in terms of either growth or delayed mortality[15,16]. However, a majority of differentially expressed passenger genes are not the causal factor and do not contribute to the overall initiation or progression of cancer, and their increase/decrease in gene expression is relatively co-incidental[15–18]. The conventional approach of designating deferentially expressed genes as the causative factor for complex human disease conditions raised several questions and doubts over the methodology[18].

Microarray Network games have the potential to accurately depict the interactions among genes as their founding premise is to consider the interactions among players governed by a network structure[19–21]. In this study, the novel Game-Theoretic-Link Relevance Index (LRI) methodology[21] is studied and applied to analyze the underlying salient genes and the associate functional annotations in CRC gene expression datasets. The resulting salient genes have an excellent potential in diapeutics, exhibiting characteristics essential for both diagnosis and therapeutics to mitigate this global complex human health condition. This work presents an opportunity to explore these salient genes further by experimental, preclinical, and clinical investigation to establish these as biomarkers.

## Materials and methods

**Colorectal cancer (CRC) patients dataset.** For this study we used meta-dataset (E-MTAB-6698) from Arrayexpress database which comprises of 15 independent GEO datasets (GSE13067, GSE13294, GSE14333, GSE15960, GSE17536, GSE17537, GSE18088, GSE18105, GSE20916, GSE23878, GSE26682, GSE33113, GSE4107, GSE4183, GSE9348). All the GEO datasets of this meta-dataset were built using a common platform (GPL570; Affymetrix Human Genome U133 Plus 2.0 Array) to prevent deviations across different platforms. In brief, a total of 1566 underlying colorectal tissue samples microarray datasets, from tumor-free (control) and primary tumors (case), were preprocessed with RMA normalization, merged and ComBat-correction for batch effect correction. This large (meta-) dataset offers very high classifying accuracy (0.997) to test on TCGA (The Cancer Genome Atlas) dataset[22] and serves as unparalleled cohort data for discovering salient genes crucial for disease phenotype development[23].

**Estimation of differentially expressed genes (DEG) in the dataset.** We assessed the microarray dataset using the conventional method of analysis. The 'limma' package[24] from Bioconductor[25] in R environment[26,27] was utilized to identify DEGs as genes associated with CRC on a pre-normalized dataset (E-MTAB-6698). Linear models were applied and empirical Bayes statistics were calculated to assess DEGs between CRC samples and healthy control samples as defined by the designed experiments[28]. The genes with Benjamini–Hochberg False Discovery Rate (FDR) controlled adjusted p-value of $\leq 0.05$ with Log2 Fold Change (LFC) $\geq 2$ (two) were considered as DEG in CRC dataset[29].

**Game-theoretic Link relevance Index (LRI) for co-expression network analysis.** Herein, we utilized the CRC dataset and evaluated each gene using the LRI method[21], as detailed in this section.

Let $(N, g^E)$ be a gene co-expression network where $N$ represents a set of genes and $g^E$ be the set of links with respect to the Microarray Experiment Situation (MES) $E =< N; S_D; S_R; A^{S_D}; A^{S_R} >$. Herein, $S_D$ and $S_R$ be the sets of samples from diseased and normal tissues, $A^{S_D}$ and $A^{S_R}$ be their expression matrices respectively. The link between $i$ and $j$ are in $g^E$ if two genes in $N$ are co-expressed. The set of all links or edges, $g^N = \{ij : i, j \in N, i \neq j\}$ is called the complete network. Let $G = \{g : g \subseteq g^N\}$ denote the set of all possible networks on $N$. Let $N(g^E)$ be the set of players who have at least one link in $g^E$ i.e. $N(g^E) = \{i : ij \in g^E for\ j \in N\}$ and $n(g^E) = |N(g^E)|$ denote the number of players involved in $g^E$. Given a $g^E \in G$, define the star of gene $i$, denoted by $g_i^E$, the set of links in $g^E$ that gene $i$ is involved in i.e.$g_i^E = \{ij : ij \subseteq g^E\ for\ j \in N(g^E)\}$. Degree of the node $i$ is expressed as $|g_i^E| = d_i(g^E)$. Microarray network game $(g, v, g^E)$ was defined with the characteristic function $v : G \rightarrow R$ which assigns a worth to each set of link $g$ representing the overall magnitude of the interaction between the genes. It follows that $v$ determines the collective influence of a set of genes connected through a network based on their

co-expression. It also follows that an equivalent form of the value function $v$ as a sum of unanimity games in a microarray network game $(g, v, g^E)$ is given by:

$$v(g) = \frac{1}{n(g^E)} \sum_{i \in N(g^E)} u_{g_i^E}(g) \tag{1}$$

where the unanimity game $u_{g_i^E}$ is defined as:

$$u_{g_i^E}(g) = \begin{cases} 1 & g_i^E \subseteq g \\ 0 & otherwise \end{cases} \tag{2}$$

The value function $v$ specifies the total value that is generated by a given network structure. The class of microarray network games with player set $N$ is denoted by $M^N$. The value function $v$ of the microarray network game $(g, v, g^E)$ picks up the information that can be used to define the role of each link in each co-expression of genes by applying suitable solution concepts of network games.

LRI allocates the total worth of the network among the genes. The allocation rule $F : G \times M^N \to R^n$ on the class of microarray network games $(g, v, g^E)$ is defined as:

$$F_i(g, v, g^E) = \frac{1}{n(g^E)} \sum_{\substack{i \in N(g') \\ g' \subseteq g}} \frac{|g_i'|}{2|g'|} \overline{\alpha}_{g'}(v)$$

Here $\overline{\alpha}_{g'}(v) = |\{i \in N(g^E) : g_i^E = g\}|$. Thus if we take $g' = g_j^E$ then $\overline{\alpha}_{g'}(v) = 1$

$$F_i(g, v, g^E) = \frac{1}{n(g^E)} \sum_{j \in N(g_i^E)} \frac{|g_i'|}{2|g_j^E|} \tag{3}$$

where $|g_i'| = |\{k \in N(g_j^E) : ik \in g_j^E\}|$, represents the number of genes associated with gene $i$ i.e. the neighbourhood of gene i in $g_j^E$ and each gene $i \in N$ receives half of the Shapley value of link $g_i'$.

An equivalent form of the LRI is:

$$F_i(g^E, v, g^E) = \frac{1}{2n(g^E)} \left( 1 + \sum_{j \in N_i(g^E)} \frac{1}{n_j(g^E)} \right) \tag{4}$$

where $N_i(g^E) = N(g_i^E) \backslash \{i\}$ and $n_j(g^E) = n(g_j^E) - 1$. $N_i(g^E)$ denotes the set of neighbours of gene i in $g^E$ and $n_j(g^E)$ the numbers of neighbours of gene j.

LRI is a unique allocation rule which satisfies four desired properties, viz.,

- **anonymity** i.e. if $v(g_1) = v(g_2)$ for all sub networks $g_1, g_2 \subseteq g$ such that they have same number of links i.e. $|g_1| = |g_2|$ then there exists $\alpha_i \in R$ for each $i \in N$ such that $F_i(g, v, g^E) = \alpha_i |g_i|$ for each link of microarray network game $(v, g^E) \in M^N$,
- the **superfluous link property** i.e. $F(g, v, g^E) = F(g \backslash ij, v, g^E)$ for all microarray network games $(v, g^E) \in M^N$ and all links $ij$ that are superfluous in $(v, g^E)$ i.e. those link which are not in $g^E$,
- **efficiency** which implies that $\sum_{i \in N} F_i(g, v, g^E) = v(g)$ for all network games $(N,v)$ i.e. the sum of the relevance of all genes should be equal to the value of whole network, and
- **additivity** if $F(g, v_1 + v_2) = F(g, v_1) + F(g, v_2)$, for each pair $(N, v_1), (N, v_2)$ of network games with component additive value functions $v_1$ and $v_2$.

For example, consider the Microarray Experiment Situation, $E =< N; S_D; S_R; A^{S_D}; A^{S_R} >$. $N = \{1, 2, 3, 4\}$ be the set of genes and $g^E = \{12, 13, 14, 23\}$ be the network on N. The value function $v$ is such that

$$v(g) = \frac{1}{4} \{u_{\{12,13,14\}} + u_{\{12,23\}} + u_{\{13,23\}} + u_{\{14\}}\}(g) \tag{5}$$

Which confers

$$v(g) = \begin{cases} 0 & if \quad g = \{12\}, \{13\}, \{23\}, \{12, 13\} \\ \frac{1}{4} & if \quad g = \{14\}, \{12, 14\}, \{12, 23\}, \{13, 14\}, \{13, 23\}, \{14, 23\} \\ \frac{1}{2} & if \quad g = \{12, 13, 14\}, \{12, 13, 23\}, \{12, 14, 23\}, \{13, 14, 23\} \\ 1 & if \quad g = g^E \end{cases} \tag{6}$$

After calculation (using Eq. 1.3), the LRI of each gene for the microarray network game is $F(g, v, g^E) = (\frac{18}{48}, \frac{11}{48}, \frac{11}{48}, \frac{8}{48})$.

**Identification of the salient genes associated with CRC.** We created an in-house script for calculating the LRI to the 4th decimal point. The genome-wide expression dataset with genes in rows and samples in columns was provided as the input matrix. With a large meta-dataset as input and the downstream analysis of

the results, the script was executed on the AMD EPYC server with an AMD7301 processor and 256 GB memory. The resulting 126 salient genes with LRI values greater than zero were considered for downstream analysis. The salient genes were evaluated for distribution statistics and compared against all the known unique cancer biomarkers from the CellMarker database[30] to ascertain the study's uniqueness and novelty. The salient LRI genes were also compared against CRC DEG to ascertain the novelty in using LRI for salient gene discovery.

**Functional analysis of salient genes for CRC: biological network construction and enrichment analysis.** To comprehend various biological roles that may be affected during the development of colon cancer, we tried to identify the ontologies from lists of 126 salient genes that were overrepresented. To avoid any changes in interpretations due to the evolution of the Gene Ontology and its annotations[31,32], we updated the reference ontology library to the latest version. Overrepresentation of the Gene Ontology (GO) terms viz. biological process (BP), molecular function (MF), and cellular component were analyzed[33–35]. Right-sided Hypergeometric (enrichment) test with a cutoff p-value at 0.05, Benjamini–Hochberg p-value corrections, Kappa Score of 0.4 and a minimum of three (3) genes per cluster threshold was set to ascertain the enrichment.

We also checked the enrichment of 126 LRI genes in terms of location on the Chromosome to verify any biased expression of a particular locus/chromosome. The enrichment was performed for Chromosome location with 2025 terms/pathways with 61570 available unique genes (with the latest updated data of 17.02.2020) as reference data. Right-sided Hypergeometric (enrichment) test with a cutoff p-value at 0.05, Benjamini–Hochberg p-value corrections, Kappa Score of 0.4 and minimum of three (3) genes per cluster threshold was set to ascertain the enrichment.

To further evaluate the role of 126 salient genes in terms of the affected biochemical processes and identification of the critical reactions and pathways, enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Reactome pathways and Reactome reactions (database updated latest on 08.05.2020) was performed with right-sided Hypergeometric (enrichment) test threshold p-value set at 0.05, Kappa Score of 0.4 and Benjamini–Hochberg p-value corrections. The obtained networks of the enriched biochemical pathways and reactions contain a variety of functional nodes and edges. All the functional enrichment analysis and visualization of the omics information was carried out as per the recommendation for standardization of the methodology[33–36].

**Protein–protein interaction amongst the salient genes.** To better visualize the role of 126 salient genes in providing functionality to a particular phenotype by interaction amongst them, we evaluate the extent of protein–protein interaction (PPI) among those genes. All the salient protein-coding genes were analyzed for PPI in the STRING database with high confidence (0.700) as threshold interaction score and all active interaction sources checked[37]. The isolated nodes were removed from the final result. The obtained networks of the enriched biochemical pathways and reactions contain a variety of functional nodes and edges. Various cluster terms were also evaluated, with Benjamini–Hochberg False Discovery Rate (FDR) less than 0.05 to identify the most significant PPI cluster.

Exploratory network analysis and statistics were evaluated using Cytoscape[33] and R[27].

**Survival analysis of the salient genes for CRC in TCGA data.** Survival analysis of the protein-coding genes in CRC patients was evaluated using The Human Protein Atlas tool[38,39]. The CRC datasets available in the webserver contain mRNA expression levels of human genes from TCGA[40–42]of 597 cancer tissue samples from persons belonging to the alive/dead and male/female sub-group. The effect of the top 10 salient protein-coding genes on these samples was investigated for the overall survival endpoint. Herein, the expression values in FPKM of individual genes in different samples with their clinical outcomes are grouped into lower higher expressions based on median expression value. Log-rank test for Kaplan–Meier plot was utilized to assess these two groups for survival endpoints.

The effects of expression of the top 10 salient genes on patients' survival across multiple CRC microarray datasets were retrieved from PrognoScan server[43] to compare, assess and comprehend the novelty in the survival analysis of the genes in TCGA data[38,39].

## Results and discussion

Biomarkers and targeted therapeutics were introduced for the early detection and clinical management of all cancer types, including CRC[7,44,45]. Therapeutics and cure of CRC include targeted medication in early diagnosis and chemotherapy and surgical resection in severe cases of metastases to other organs and tissues followed by medications[6,8–11]. Yet, recurrence of CRC in the presence of poor diagnostic measures was reportedly found to cause additional risk and reduce the life expectance of the people[2,7,9,10,44,46] which can be attributed to widespread occurrences and recurrence of CRC, thus, making it one of the most dreaded diseases in the world[1–3,5,6,46]. Significant challenges were evident in successfully implementing specific biomarkers as a tool for cancer diapeutics[7,47,48]. Furthermore, despite several advances in cancer diapeutics, CRC continues to remain an unabated disease eventually leading to the death of the patient[1,2,5,6,49]. Therefore, a retrospection on our present knowledge on the factors with a prime etiological role in CRC is a must for mitigating the occurrence of CRC through targeted diapeutics.

Conventional algorithms for discovering genes of importance/biomarkers responsible for a physiological condition such as cancer rely on differentially expressed genes considered to be critical factors in the progression or manifestation of cancer condition. The conventional method identifies and prioritizes genes based on the degree of difference in expression values in cases (cancer) compared to control (normal healthy). These differentially expressed genes exhibit a high fold difference between gene expression values in cancer cases compared to normal conditions. In other words, the conventional methods convey that the genes with a greater degree of difference in expression level in disease samples than normal samples are more important than genes with a lesser degree of

difference[18]. These methods possess many immediate challenges as the method dictates that the genes that exhibit greater fold difference in gene expression values are considered of greater importance, which may not be valid[18].

Genes that initiate the cancer progression might have less fold difference in expression values than downstream effector genes, which often exhibit higher fold difference in expression[15–18]. Many passenger genes may exhibit greater fold difference though their contribution in the manifestation of the cancer is incidental[15–17]. On the other hand, driver genes, which are the causal factor, with comparatively lower fold difference in gene expression contribute more in the progression and advancement of the cancerous cell[15,16]. Also, these methods ignore the contribution of each gene in the overall gene network of cancer/case. Reassessment of diagnostic and prognostic markers for breast cancer and other cancer type were reported previously[18,50]. The investigators questioned the conventional approach and revealed that designating an etiological role in complex human disease conditions simply to the higher expressed genes may not be the correct methodology[18].

Game theory (GT) has unlocked newer frontiers in solving various bioinformatics and computational biology challenges, from evolutionary genetics and virulence evolution modelling to high-throughput genomics data and biological networks[19,19–21,51–56]. Coalition GT on large-scale biological networks bestows estimation of the power of each gene governing biological pathways of interest and the associated etiological role in complex human health conditions[19,20]. GT application in quantitative evaluation of prominence of genes, by considering their relationships with others, in initiation and progression of disease condition contributed immensely in understanding the behaviors of salient genes in manifesting disease[54,55]. Cooperative Game theoretical approaches such as Shapley value and Banzhaf value provided valuable insight into gene expression data analysis by screening the dataset for the most relevant genes involved in the condition of interest[52,53]. Previously, the GT approach exhibited its ability at par with classical centrality indices in evaluating each gene by its relevance. It also emphasized the function of genes as nodes present in the periphery of a co-expression network in modulating the complex biological pathways[56].

We adopted a Game Theoretic approach in this model, especially the approach of Network games. An improved GT method, LRI, was recently proposed to identify salient genes involved in cancer or other metabolic syndromes[21]. The LRI in this model is brought from the concept of Shapley value of cooperative game theory in networks which can be used as a relevant approach for the classification of genes[21]. A substantial attribute of LRI model of game theory is that it provides an innovative property-driven classification of the use of Shapley value as an index to validate and contextualize genes[57,58]. In microarray games, Shapley value was used to quantitatively evaluate the underlying genes involved in disease manifestation and characterise their role in gene-regulatory pathways[54,56,59]. LRI, on the other hand, utilizes a co-operative framework to analyze the microarray data of gene co-expression networks where genes and their connecting links play a significant role in determining the overall structure. It emphasizes that when we consider such a co-expression network, LRI can substitute Shapley value. This is because LRI focuses on the linking abilities of the genes as a suitable candidate to demonstrate the significance of the genes and is based on the position value (a link based allocation rule)[21], while Shapley value is based on the Myerson value, which is a player based allocation rule[54,55]. LRI establishes that any network game can precisely describe the gene interactions as it considers the cooperation among genes and how the genes are connected to a network providing a comprehensive description of the genetic markers and their combined effects[21].

E-MTAB-6698 is a large (meta-) dataset that comprises gene expression of colorectal tissue samples data with relevant clinical history and conditions of 1566 persons from both the biological gender. The whole-genome gene expression microarray data built on the GPL570 platform includes 121 colon samples from normal persons (as control), 1393 samples from the diseased part of the CRC patient, 37 samples from Adenoma patients, and 15 samples from patients suffering from Inflammatory bowel diseases. The overall dataset already proved its effectiveness by offering a very high degree of classification accuracy (0.997) to test the RNAseq dataset during training and modelling the disease condition. It functions as an unmatched cohort data for investigating and determining salient genes crucial for CRC development[22,23].

Herein, we applied this game-theoretic LRI scoring[21] on the high-throughput CRC transcriptomics dataset to identify salient genes in CRC. Contrary to the conventional approach, the Game-Theoretic Link relevance Index identifies a gene's importance by considering genes' contribution to overall disease manifestation.

We obtained 126 genes with a positive LRI score (LRI > 0) (refer to Supplementary File Table S1) which we referred as salient genes in the article. These 126 salient genes consist of 117 protein-coding genes, 8 non-coding RNA and 1 uncharacterized gene. Of these genes, four (4) were mapped on the X chromosome and the rest one hundred and twenty-two (122) on autosomal chromosomes. None of these 126 genes was mapped on the Y chromosome. Of these 126, the top 15 genes with the highest LRI score (Table 1) consist of one ncRNA and 14 protein-coding genes. *MIR143HG* and *AMOTL1* scored the highest LRI score (0.01604), followed by *ACTG2* (0.01587).

The distribution of the LRI value of 126 salient genes was analyzed to understand the nature of the data. Other genes with zero (0) LRI scores were not considered here for distribution study. A violin plot (Fig. 1A) depicts distributions of LRI values of 126 genes using density curves where the width of the curve specifies the approximate frequency of data points in that region. Quantile–quantile (Q–Q) plot (Fig. 1B) exhibits LRI data points falling in the middle of the plot and curve off in the extremities, indicating that the behaviors of the LRI data points are suggestive of high extreme values than would be expected if the data were normally distributed. The density-cum-histogram plot (Fig. 1C) describes the distribution of the LRI values of 126 genes against the count of genes (or proportion in secondary axis). All the exploratory data distribution analysis suggests that the distribution of LRI values is not normal and that disproportionate extremes of LRI values are assigned to those 126 genes.

CellMarker is an enormous curated database of biomarkers, especially at the single-cell level containing more than 22,000 cell markers of different cell types, including cancer cells[30]. To assess the uniqueness and novelty of

| S. no | LRI score | Gene ID | Type of gene | Full name from nomenclature authority |
|---|---|---|---|---|
| 1 | 0.016045 | *MIR143HG* | ncRNA | Cardiac mesoderm enhancer—associated non—coding RNA |
| 2 | 0.016045 | *AMOTL1* | Protein-coding | Angiomotin like 1 |
| 3 | 0.015873 | *ACTG2* | Protein-coding | Actin gamma 2, smooth muscle |
| 4 | 0.011054 | *FILIP1* | Protein-coding | Filamin A interacting protein 1 |
| 5 | 0.011054 | *ARHGEF17* | Protein-coding | Rho guanine nucleotide exchange factor 17 |
| 6 | 0.011054 | *FAM219B* | Protein-coding | Family with sequence similarity 219 member B |
| 7 | 0.00959 | *ITPKB* | Protein-coding | Inositol—trisphosphate 3-kinase B |
| 8 | 0.00959 | *TOP2A* | Protein-coding | DNA topoisomerase II alpha |
| 9 | 0.00959 | *HAND1* | Protein-coding | Heart and neural crest derivatives expressed 1 |
| 10 | 0.00959 | *SERINC2* | Protein-coding | Serine incorporator 2 |
| 11 | 0.009081 | *TRAP1* | Protein-coding | TNF receptor associated protein 1 |
| 12 | 0.009081 | *CAMSAP1* | Protein-coding | Calmodulin regulated spectrin cassociated protein 1 |
| 13 | 0.009081 | *APOBR* | Protein-coding | Apolipoprotein B receptor |
| 14 | 0.009081 | *PAG1* | Protein-coding | Phosphoprotein membrane anchor with glycosphingolipid microdomains 1 |
| 15 | 0.009081 | *MRPS9* | Protein-coding | Mitochondrial ribosomal protein S9 |

**Table 1.** LRI score of top 15 salient genes with their types and full name from nomenclature authority.

the result in the present investigation, we extracted information of all the known biomarkers from the CellMarker database. Information of all the markers genes for cell types including Cancer cell, Cancer stem cell, Cancer stem-like cell, Tumor endothelial cell, and Tumor-propagating cell from the CellMarker database was retrieved. All the cell type individually provided information of 180 genes, which were further reduced to 171 after removing duplicates. We mapped 126 salient genes to Entrez ID for maintaining uniformity. One hundred twenty-four genes mapped to their corresponding Entrez ID except for two Genes IDs viz, *LOC100129461* and *LOC400965*. A comparative analysis (Fig. 1D) revealed that two genes, viz, *ITGA5*, and *MCAM* exhibited overlapped with the known biomarkers, suggesting that the information about these two genes is already present in the existing curated knowledge base cancer biomarkers. Except for these two, however, all the salient genes demonstrated no overlap with cancer biomarkers suggesting the resulting salient genes information exhibits novelty.

The conventional method of identifying genes associated with disease relies on the assumption that the greater the difference between the expression of the gene under the normal sample and the disease sample, the greater the chance that the gene is responsible for disease occurrence. The conventional method identifies the genes associated with CRC disease by isolating genes differentially expressed in the CRC sample compared to normal. The DEGs of CRC (Table 2) were compared to salient genes to ascertain the novelty in the LRI approach. The analysis (Fig. 1E) revealed that only two salient genes viz, *MT1M* (LRI value 0.007937), and *SI* (LRI value 0.007937) exhibited overlapped with the DEGs suggesting that the genes identified using the LRI approach is very much different from the conventional approach. These non-overlapping salient genes that do not overlap with the known biomarkers or with DEGs (Fig. 1D,E) present an opportunity to assess their role as diapeutics biomarkers further.

These non-overlapping genes present an opportunity to assess them further for their role as diapeutics biomarkers.

The 126 salient genes were found to be associated with eleven Biological Process terms falling in seven GO groups (Fig. 2; Supplementary File Table S2), exhibiting overrepresentation. The enriched BP terms include ryanodine-sensitive calcium-release channel activity (GO:0005219), wound healing, spreading of cells (GO:0044319), muscle tissue morphogenesis (GO:0060415), platelet aggregation (GO:0070527), mesenchyme morphogenesis (GO:0072132), regulation of muscle contraction (GO:0006937), regulation of cardiac muscle contraction (GO:0055117), positive regulation of blood circulation (GO:1903524), positive regulation of muscle contraction (GO:0045933), negative regulation of vascular smooth muscle cell proliferation (GO:1904706) and regulation of smooth muscle contraction (GO:0006940). Ryanodine-sensitive calcium-release channel activity results in several skeletal myopathies due to dysregulation of intracellular $Ca^{2+}$ and several muscle myopathies[60]. Wound healing and spreading of cells process is marked by collective migration of epithelial cells in the form of coherent sheets to heal wounds[61,62]. During cancer, one of the most prevalent phenomena is muscle dysfunction, where patients, irrespective of tumor stage and nutritional state, are subjected to compromised muscular function[63]. There have been several evidence that mitochondrial dysfunction can be induced by chemotherapy, which in turn contributes to muscle atrophy[64–68]. The biological process of tumor–induced platelet aggregation has several mechanisms involved which vary from tumor cell to the other and are generally activated by the generation of tumor cell-induced thrombin[69].

Furthermore, *AHNAK*, *CASQ2*, *CCL2*, *CHRM3*, *FXYD6*, *PLN*, and *PRKCE* genes are associated with the GO term for regulation of ion transmembrane transporter activity (GO:0032412) exhibited overrepresentation of the MF (Fig. 3A; Supplementary File Table S3). These genes may affect the progression of CRC as a consequence of perturbation of the critical process that modulates the activity of an ion transporter. This GO term demonstrated overrepresentation in a list of Cytosine—phosphate—Guanine (CpG) sites that exhibited a steady depolarization change[70].

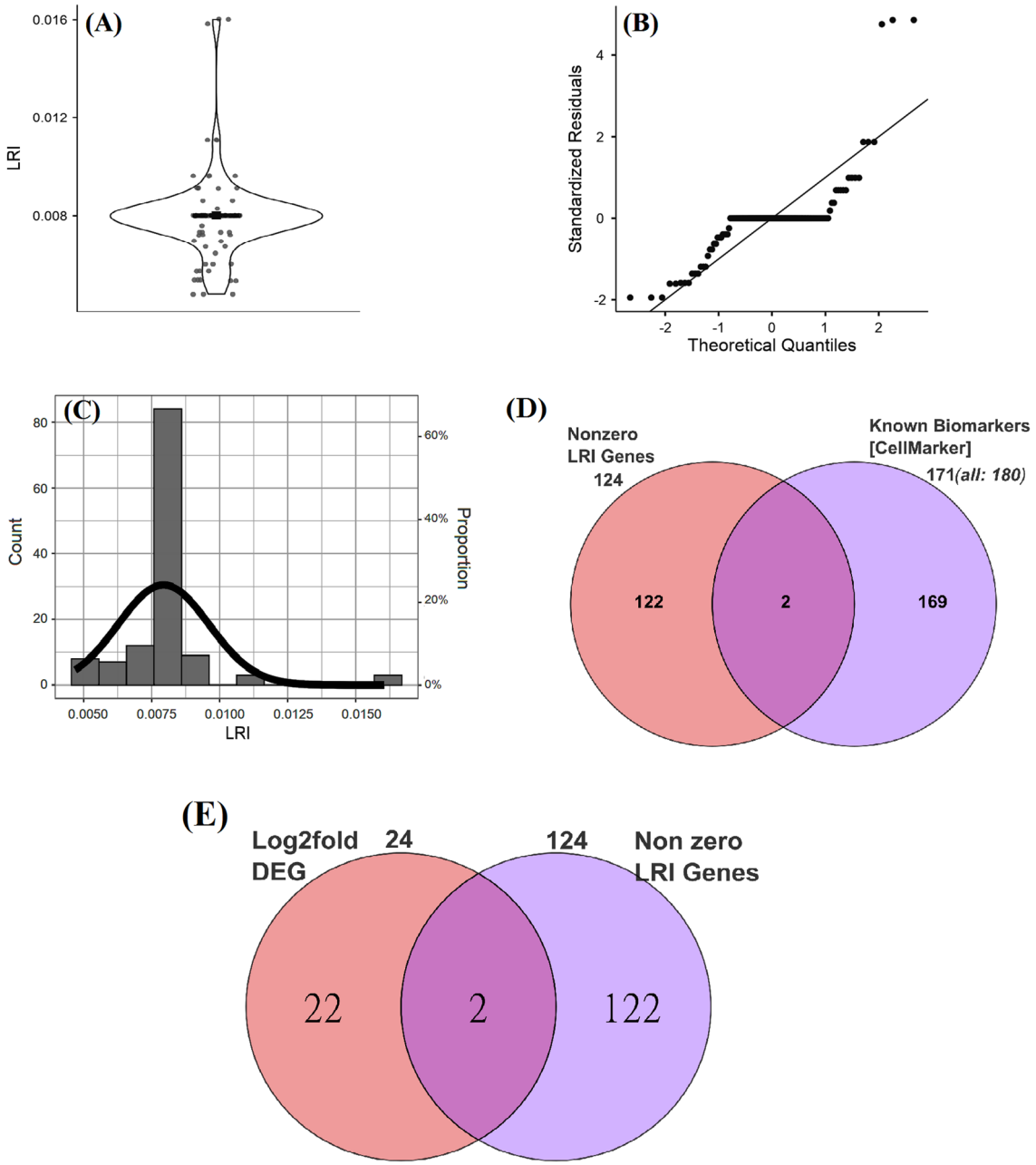


**Figure 1.** Distribution of LRI values of the salient genes. The figure exhibits the distribution of 126 salient genes. Violin plot (**A**) with jittered points data points and mean value, Q–Q plot (**B**) and histogram combined density plot (**C**) to show distribution LRI values of these 126 salient genes. (**D**) Comparison of the 124 salient genes (out of 126 Gene IDs, two Gene IDs did not map to Entrez ID) with 171 (after removing duplicates from the total 180 genes) unique cancer biomarkers from CellMarker database exhibits overlap of only two genes and 122 genes exhibits uniqueness. (**E**) Comparison of the 124 salient genes with 24 DEGs of CRC exhibits overlap of only two genes.

Overrepresented CC include Sarcoplasmic reticulum membrane (GO:0033017) with three associated genes (*CASQ2*, *PLN*, and *RYR3*), myofibril (GO:0030016) with nine (*ACTC1*, *AHNAK*, *CASQ2*, *FLNA*, *LMOD1*, *MYL9*, *RYR3*, *TPM1*, and *VCL*), sarcomere (GO:0030017) with seven (*ACTC1*, *CASQ2*, *FLNA*, *LMOD1*, *MYL9*, *RYR3*, *TPM1*), and I band (GO:0031674) with five associated genes (*ACTC1*, *CASQ2*, *FLNA*, *MYL9*, *RYR3*) (Fig. 3B; Supplementary File Table S4). *CASQ2* and *RYR3* are common genes between the Sarcoplasmic reticulum membrane (GO:0033017) and myofibril (GO:0030016) ontology terms. The Sarcoplasmic reticulum membrane has been associated with inherited dysfunctions and deficiencies like cardiac arrhythmias[71]. The enzymes involved in Sarco/endoplasmic reticulum calcium transport ATPases play a crucial role in loss or reduction of colon carcinomas and apoptosis[72]. Different myofibrils have been found to be associated with either oncogenic or tumor suppressor roles in different cancers like lung cancer, breast cancer, prostate and CRC[73].

| Genes | LFC | Average expression | t value | p-value | FDR adjusted p-value | B value |
|---|---|---|---|---|---|---|
| *GCG* | −2.38358 | 4.969389 | −15.752 | 6.92E−52 | 1.29E−48 | 107.0804 |
| *FOXQ1* | 2.41793 | 9.809112 | 15.58703 | 6.45E−51 | 9.47E−48 | 104.8698 |
| *CA1* | −2.68165 | 7.905425 | −15.5739 | 7.71E−51 | 1.06E−47 | 104.6939 |
| *CEMIP* | 2.019388 | 8.943608 | 15.43317 | 5.10E−50 | 6.17E−47 | 102.8227 |
| *CLDN8* | −2.37615 | 4.975604 | −14.9159 | 4.78E−47 | 3.39E−44 | 96.04866 |
| *AQP8* | −2.20027 | 6.339491 | −14.6653 | 1.24E−45 | 6.68E−43 | 92.8299 |
| *SLC4A4* | −2.69512 | 6.884806 | −14.5002 | 1.03E−44 | 4.82E−42 | 90.73028 |
| *PKIB* | −2.09604 | 7.226978 | −13.9028 | 1.91E−41 | 5.58E−39 | 83.28646 |
| *MT1M** | *−2.12435* | *6.256335* | *−13.8111* | *5.93E−41* | *1.54E−38* | *82.16609* |
| *MS4A12* | −2.71223 | 6.861364 | −13.7431 | 1.37E−40 | 3.35E−38 | 81.33784 |
| *ZG16* | −2.74676 | 8.079408 | −13.6444 | 4.59E−40 | 1.03E−37 | 80.14146 |
| *CLCA4* | −3.05432 | 6.704506 | −13.5764 | 1.05E−39 | 2.14E−37 | 79.32192 |
| *CA2* | −2.43465 | 8.684428 | −12.687 | 3.99E−35 | 4.53E−33 | 68.89684 |
| *ADH1C* | −2.11602 | 7.314942 | −11.9071 | 2.59E−31 | 1.71E−29 | 60.22568 |
| *SI** | *−2.37608* | *6.102779* | *−11.4662* | *3.02E−29* | *1.51E−27* | *55.52607* |
| *PCK1* | −2.02539 | 8.368289 | −11.4555 | 3.38E−29 | 1.69E−27 | 55.41363 |
| *MMP3* | 2.000444 | 8.471801 | 11.2982 | 1.78E−28 | 8.17E−27 | 53.77398 |
| *MMP1* | 2.123265 | 8.813216 | 10.83644 | 2.09E−26 | 7.03E−25 | 49.07185 |
| *KRT23* | 2.006019 | 7.476861 | 10.55127 | 3.65E−25 | 1.08E−23 | 46.25175 |
| *HEPACAM2* | −2.06678 | 6.938537 | −10.292 | 4.65E−24 | 1.23E−22 | 43.74442 |
| *CEACAM7* | −2.10091 | 9.917928 | −10.0767 | 3.69E−23 | 8.83E−22 | 41.70331 |
| *SLC26A3* | −2.45108 | 9.32326 | −9.64708 | 2.06E−21 | 3.95E−20 | 37.74335 |
| *ITLN1* | −2.20724 | 7.95737 | −9.6134 | 2.80E−21 | 5.29E−20 | 37.43937 |
| *CLCA1* | −2.07261 | 8.930104 | −8.52471 | 3.65E−17 | 4.48E−16 | 28.12584 |

**Table 2.** DEGs (Differentially Expressed Genes) in the CRC dataset. The genes that exhibited adjusted p-value of ≤ 0.05 and Log2 Fold Change (LFC) ≥ 2 (two) were considered as DEG in the CRC dataset. DEGs that are also present in the list of salient genes are denoted by (*).

No enrichment with respect to chromosome location was observed for the 126 salient genes. These salient genes were observed to be distributed through the genome and not tandemly in a particular affected region. This suggests that the aberrant gene expression of the genes is not a consequence of the activation of a particular region in a chromosome.

Thirty eight (38) KEGG pathway terms, viz., wound healing, spreading of cells (GO:0044319), acylglycerol catabolic process (GO:0046464), regulation of hair follicle development (GO:0051797), platelet aggregation (GO:0070527), mesenchyme morphogenesis (GO:0072132), positive regulation of cellular carbohydrate metabolic process (GO:0010676), ATP transmembrane transporter activity (GO:0005347), anion:anion antiporter activity (GO:0015301), positive regulation of blood circulation (GO:1903524), positive regulation of muscle contraction (GO:0045933), regulation of muscle contraction (GO:0006937), smooth muscle cell migration (GO:0014909), muscle filament sliding (GO:0030049), regulation of smooth muscle cell migration (GO:0014910), positive regulation of smooth muscle contraction (GO:0045987), negative regulation of vascular smooth muscle cell proliferation (GO:1904706), regulation of smooth muscle contraction (GO:0006940), sarcomere organization (GO:0045214), regulation of cardiac muscle contraction (GO:0055117), negative regulation of vascular associated smooth muscle cell migration (GO:1904753), Dilated cardiomyopathy (DCM) (KEGG:05414), ryanodine-sensitive calcium-release channel activity (GO:0005219), release of sequestered calcium ion into cytosol by sarcoplasmic reticulum (GO:0014808), regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion (GO:0010881), relaxation of cardiac muscle (GO:0055119), regulation of muscle contraction (GO:0006937), muscle filament sliding (GO:0030049), negative regulation of neurotransmitter uptake (GO:0051581), myofibril assembly (GO:0030239), muscle tissue morphogenesis (GO:0060415), cellular response to caffeine (GO:0071313), cardiac ventricle morphogenesis (GO:0003208), negative regulation of cation transmembrane transport (GO:1904063), sarcomere organization (GO:0045214), regulation of cardiac muscle contraction (GO:0055117), regulation of cardiac muscle cell contraction (GO:0086004), negative regulation of vascular associated smooth muscle cell migration (GO:1904753), and negative regulation of calcium ion transmembrane transporter activity (GO:1901020) exhibited significantly high enrichment (Fig. 4; Supplementary File Table S5). Acylglycerol catabolic process has been used as a biomarker for the diagnosis and/or prognosis of CRC, and the enzymes of acylglycerols are involved in CRC tumor growth survival and metastasis[74]. Changes in the cellular carbohydrate metabolic process may precede the acquisition of driver mutations, ultimately leading to colonocyte transformation. These changes may not be uniform but rely on different pathways to adapt to nutrient availability[75]. Muscular contraction was found to be enriched in the signal pathway of the differentially expressed genes associated with the early onset of CRC[29].

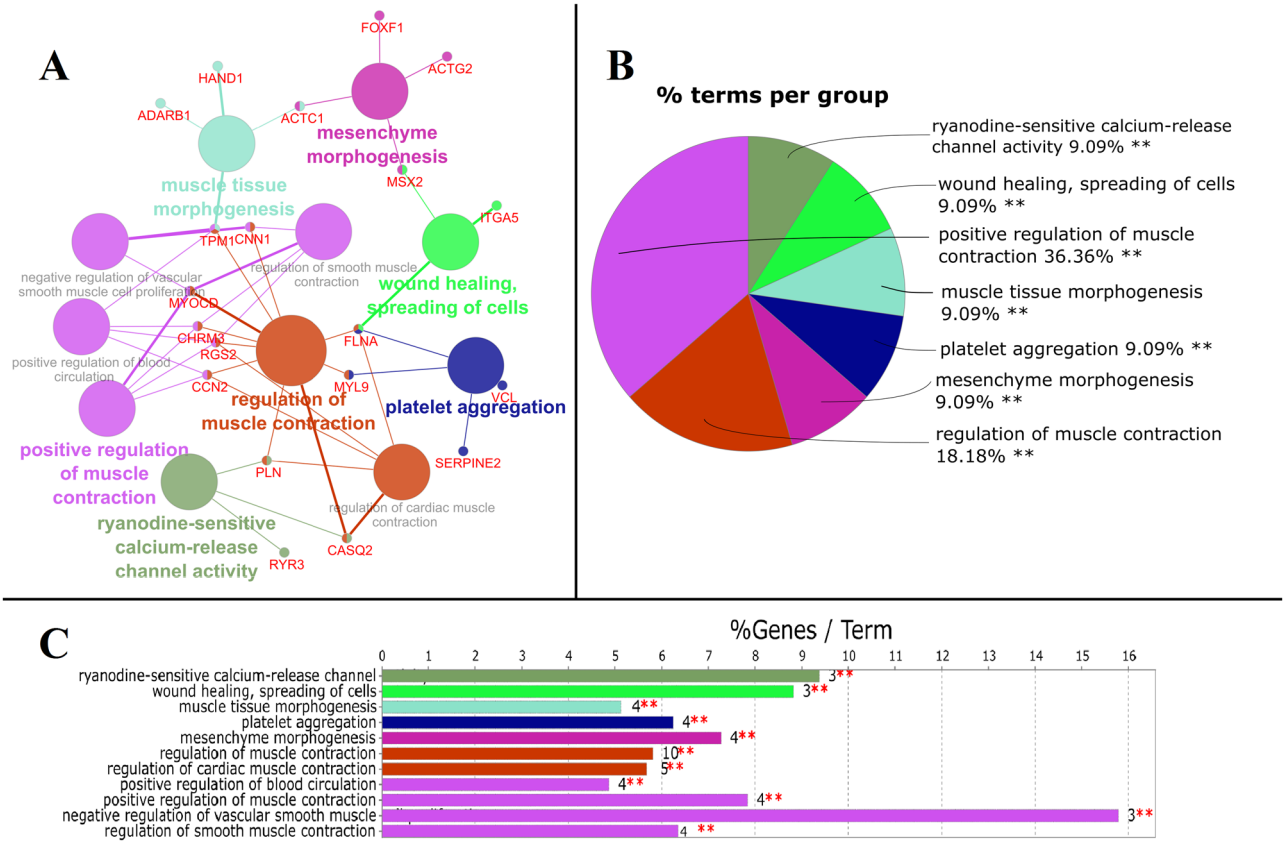


**Figure 2.** Enrichment of the major Biological Processes (BP) associated with the 126 salient genes. (**A**) Represents the network of various sub-ontologies, and associated genes, (**B**) describes percentage terms per group for various BP that are significantly enriched in pie chart and (**C**) number of genes in each term with significance sign. Node size is inversely proportioned to the *p*-value, i.e., the lower the value, the bigger the node size and color represent a different group of terms. *Significant at $p \leq 0.05$, and **Significant at $p \leq 0.001$.

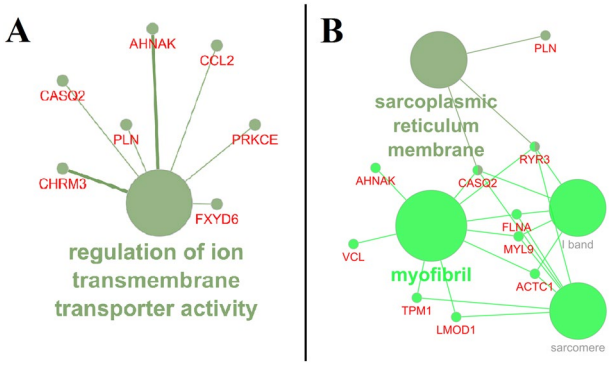


**Figure 3.** Enrichment of the major Molecular Function (MF) (**A**) and Cellular component (CC) associated with the 126 salient genes. (**A**) Describes percentage terms per group for various MF that are significantly enriched, and (**B**) shows various sub-ontologies of CC and associated genes. Node size is inversely proportional to the p-value, i.e., the smaller the value more considerable the node size and colour represents a different group of terms.

We searched for enriched Reactome pathways using all available genes as a reference from the database. Three terms viz. Muscle contraction, Smooth Muscle Contraction, Ion homeostasis exhibited high significant enrichment (Fig. 5A; Supplementary File Table S6). All the three Reactome pathway terms revealed equal enrichment with 33.33% of the total terms distributed per group. Five (5) genes, viz. *ACTG2*, *LMOD1*, *MYL9*, *TPM1*, *VCL* are associated with Smooth Muscle Contraction (R-HSA:445355), four (4) genes viz. *CASQ2*, *FXYD6*, *PLN*, *RYR3* are associated with Ion homeostasis (R-HSA:5578775) and ten (10) genes, viz, *ACTC1*, *ACTG2*, *CASQ2*, *FXYD6*,

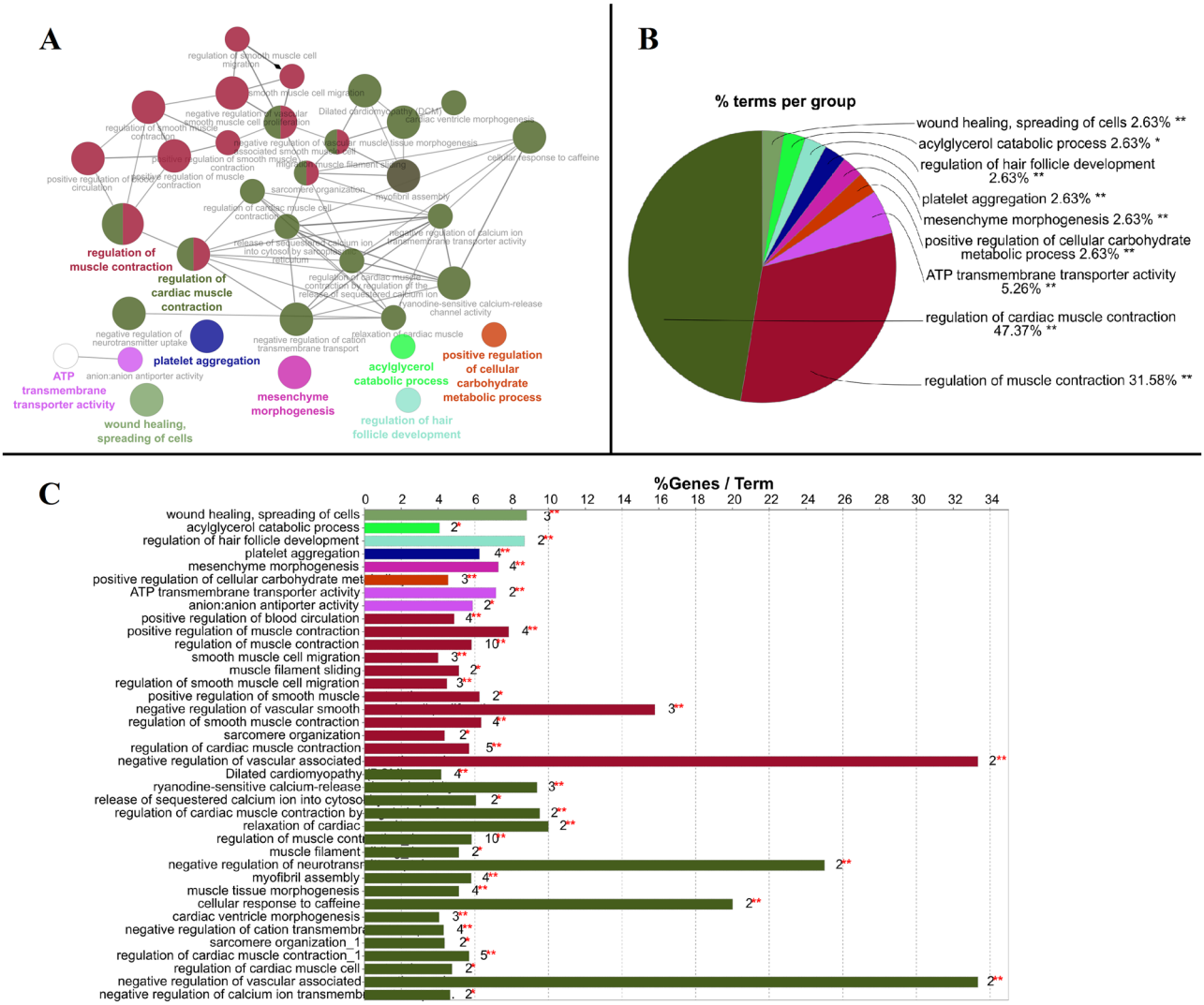


**Figure 4.** Enrichment of the major KEGG pathways associated with the 126 LRI genes. (**A**) Represents the network of various pathways and sub-pathways and associated genes where the size of the node is inversely proportional to the p-value, i.e., the lower the p-value, the bigger the node size and colour represents a different group of pathway terms. (**B**) Describes percentage terms per group for various parent pathways significantly enriched in pie chart, and (**C**) describes the number of genes in each pathway and sub-pathway term with significance sign. *Significant at $p \leq 0.05$, and **Significant at $p \leq 0.001$.

*LMOD1*, *MYL9*, *PLN*, *RYR3*, *TPM1*, *VCL*, are associated with Muscle contraction (R-HSA:397014). Of the 126 salient genes of CRC, these genes are primarily associated with muscle contraction. Their aberrant expression must affect the associated processes and functional proteins in the progression of CRC. Muscle contraction and dysfunction have been found to be intensely associated with CRC, and their respective molecular functions indicate that they could possibly be the therapeutic targets of CRC[29,76]. Their aberrant expression must affect the associated processes and functional proteins in the progression of CRC. Ion homeostasis plays an indispensable role in the physiology of the gastrointestinal tract, and any dysregulation is an indication of gastrointestinal cancer. They can, therefore, be used as a useful prognostic biomarker for gastrointestinal cancer[77]. Cancer metastasis has often been found to be accompanied by skewing in ion homeostasis[78].

We also searched for enriched Reactome reactions using all available genes as a reference from the database. Five (5) genes viz. *LMOD1*, *TPM1*, *VCL*, *ACTG2*, and *MYL9* contributed towards very high significant enrichment (100% of the terms per group) exhibited by the term 'ATP Hydrolysis By Myosin (R-HSA:445699)' (Fig. 5B; Supplementary File Table S7). With Kappa Score of 0.4, which is used to define term-term interrelations (edges) and functional groups based on shared genes between the terms, Reactome reactions viz. ATP Hydrolysis by Myosin (R-HSA:445699), Myosin Binds ATP (R-HSA:445700), Calcium Binds Caldesmon (R-HSA:445704), Release of ADP From Myosin (R-HSA:445705) exhibited high enrichment suggesting that the salient genes majorly associated with smooth muscle contractions activity. Results obtained recently also suggest that the Muscle contraction and vascular smooth muscle contraction pathway as major affected molecular mechanisms in CRC[29], which corroborate the inference of our present work.

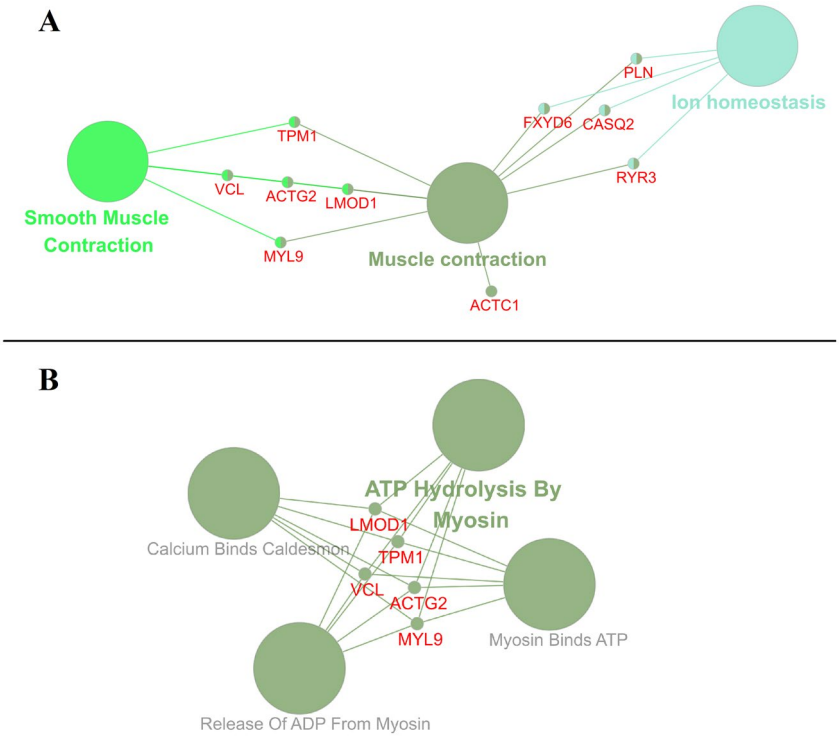


**Figure 5.** Enrichment of the major Reactome pathways and reactions. The figure represents the Enrichment of the Reactome pathways (**A**) and reactions (**B**) associated with the 126 salient genes (Labeled in red colour). Node size is inversely proportional to the p-value, i.e., the lower the p-value, the bigger the node size and colour represents a different group of terms.

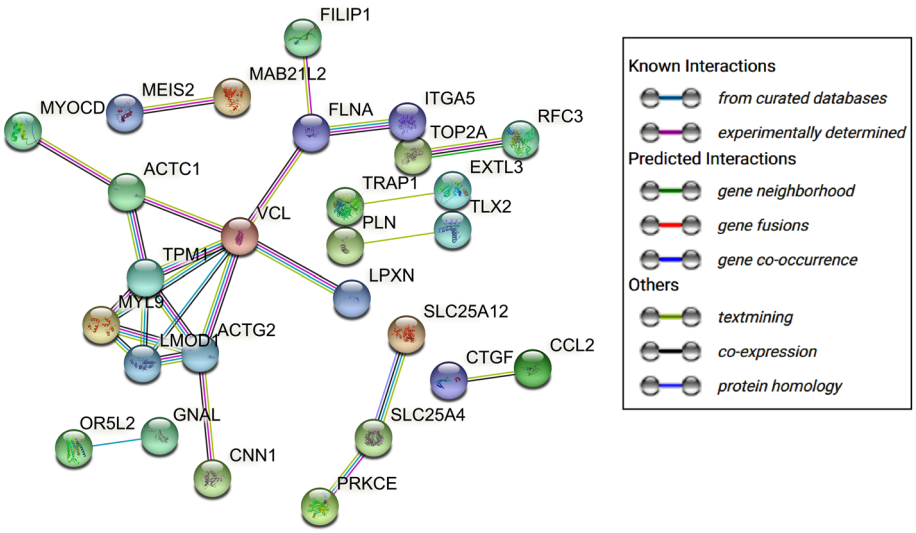


**Figure 6.** Protein–Protein interaction (PPI) network among the 126 salient genes. At a confidence score of 0.700, the figure exhibits major interactions among the protein-coding genes as node and various interactions as edges. The colored nodes are query proteins with 3D structures (if any) inside the node and edge color represents evidence as an indicator of interactions among the proteins. The isolated nodes were removed from the network.

We investigated protein–protein interaction (PPI) among the salient genes (Fig. 6). STRING database[37] was able to identify 116 protein-coding genes as nodes and presented the PPI network. At a threshold confidence score of 0.700 (high confidence score), a total of 26 edges were formed with an average node degree equal to 0.448, the average local clustering coefficient of 0.198. With expected number of edges equals 9, the network exhibited significant enrichment (p-value = 0.00000141 < 0.05). Total ten clusters were found with FDR p-value less than

0.05 suggesting these clusters are significantly enriched (Supplementary File Table S8 The most prominent clusters, CL:1326 and CL: 1328, consist of 10 and 8 protein-coding genes, respectively, and are associated with 'Muscle protein and Myofbril assembly'. The smallest clusters, CL:25786, CL:1449, CL:1577, CL:6451, consist of only two protein-coding genes each. All the 10 clusters exhibited cluster value greater than 1 for Strength which is a measure of enrichment effect and is expressed as Log10 value of the ratio between observed gene count in the network and expected gene count suggesting and value lesser than 0.05 for FDR, suggesting the results of high enrichments are statistically significant (Supplementary File Table S8).

It is also pertinent to note that some in vitro and in vivo reports suggest the top three genes viz. *MIR143HG*, *AMOTL1*, and *ACTG2* may be associated with other cancer subtypes. However, after a thorough literature survey, we were not able to mine any data suggesting their etiological role in CRC development.

MIR143HG (LRI = 0.016045) is a highly conserved long non-coding RNA that hosts miR-143/145 cluster that modulates smooth muscle cell differentiation and remodelling[79]. The association of the MiR143HG with bladder cancer and hepatocellular carcinoma is well established[80,81]. MiR143HG/MIR-1275/AXIN2 tri-axis is directly responsible for the onset of bladder cancer its progression by tempering the Wnt/β-catenin pathway. Its downregulation is directly associated with the development and progression of bladder cancer[80] while an upregulation is associated with hepatocellular carcinoma[81]. It also has an important functional role(s) in cardiovascular system development[79].

AMOTL1 (LRI = 0.016045) is a motin family member or Angiomotins (AMOTs) that dictates the functioning of several bioprocesses, including tight junction formation, angiogenesis, cell polarity, and migration[82,83]. Previous reports suggest *AMOTL1* is an oncogene and their dysregulated expression affects promotion, proliferation, migration and relapse of cancer cells, including prostate cancer, renal cell cancer, cervical cancer, liver cancer, head and neck squamous cell carcinoma, bladder cancer, and osteosarcoma[82–85]. Contrary to this, it also exhibits tumor suppression function inhibiting cancer cells' growth in glioblastoma, ovarian cancer, and lung cancer[82].

The actin gamma smooth muscle 2 (*ACTG2*) gene ((LRI = 0.015873), belonging to the actin protein family, is imperative for maintaining the cytoskeleton through the regulation of cell movement and muscle contraction[86]. Genome sequencing studies have revealed that a homozygous and a heterozygous variant of *ACTG2* is associated with gastrointestinal dysfunction[87]. Studies have demonstrated that the over-expression of *ACTG2* has been found to play a critical role in the progression of hepatocellular carcinoma[88] and bladder cancer[89]. Conversely, another finding has described a concomitant improved survival and more aggressive phenotype with a higher expression of *ACTG2*[90,91]. However, a lower expression of the gene was associated with normal colon tissue in contrast to colon carcinoma[92], while imperceptible expression levels of *ACTG2* have been associated with the metastasis of lymph nodes[93]. A recent bioinformatics work comparing samples of 12 CRC patients and 10 healthy control group also hinted at the possible role of *ACTG2* in manifesting CRC[29], which is consistent with our findings.

Filamin A interacting protein 1 (*FILIP1*), a potent antivascular cancer therapeutic, has been demonstrated previously to be a key modulator of angiogenesis's inhibitory effects[94]. Moreover, the *FILIP1* is also found to inhibit cell invasion and metastasis in ovarian cancer by downregulating the Wnt pathway[95]. On the other hand, the Rho guanine nucleotide exchange factor 1 protein (*ARHGEF17* gene), formerly known as a guanine nucleotide exchange factor (*GEF*), is a vital mitotic gene. ARHGEF17 is indispensable for the spindle assembly checkpoint and targets mitotic kinase Mps1 to mitotic kinetochores[96]. It is also presumed to be responsible for lung carcinoma cell migration stimulated with lysophosphatidic acid[97]. The Family with Sequence Similarity 219 Member B (*FAM219B*), a paralog of *FAM219A* gene, is a protein-coding gene. The diseases associated with it include Metachondromatosis and Leopard Syndrome 1[98,99].

The information on the role of the top 10 genes according to LRI in CRC development is either not present or is negligible. Our data mining suggests the unavailability of any previous information indicating the role of *FAM219B* in any cancer type. Notwithstanding the potential role of salient genes in other cancer types, the central role of *MiR143HG*, *AMOTL1*, *ACTG2*, *FILIP1*, *ARHGEF17*, *FAM219B* in the progression of CRC is inadequate and lacking in previous reports. Previous reports on *TOP2A*[100,101], *ITPKB*[101], *HAND1*[102], *SERINC2*[103–105] present a conjectural view of the importance of the gene in the progression of the CRC.

Identifying the top salient genes as diapeutics biomarkers for CRC will be critical to diagnostics, predicting the disease's occurrence/recurrence, and improvising the therapeutics. Major statistical difference in weak expression and high expression of genes in Kaplan–Meier (KM) survival analysis highlights the importance of genes with respect to their significant contribution to cancer progression and development[50]. The top 10 protein-coding genes were investigated for patients' survivability and their expression (Fig. 7). Log-rank p-value for KM plot indicates a correlation between patient's survival and gene expression level. Statistically significant results (log-rank p-value ≤ 0.05) in the overall survival endpoint suggest that the change in expression of genes compared to the cutoff threshold at the molecular level results in a notable difference in the overall patient's survival probability. The top 10 salient genes expression exhibited statistically significant results (log-rank p-value ≤ 0.05), connoting the high significance of these genes in causing CRC progression and development during perturbed expression. There exists a discernible difference in 5-year survival for patients with lower expression compared to higher expression than their respective expression cutoff for a 5-year duration. The median survival time between lower and higher expression of the genes are significantly different. Among the top 10 salient protein-coding genes, *AMOTL1*, *ACTG2*, *FILIP1*, *ARHGEF17*, *FAM219B*, *ITPKB*, *HAND1* exhibited reduced survival probability with higher expression, contrary to the rest of the other genes, viz. *TOP2A*, *TRAP1*, *SERINC2*, which exhibited reduced survival probability with lower gene expression (Fig. 7, also refer Supplementary File Figure S1–S10 for enlarged view).

All the genes exhibit a distinct difference in the survival endpoints between the two expressions. The KM plot of the top 10 protein-coding salient genes (highlighted by the LRI score) revealed a significant contribution to the CRC patients' overall outcome. A survey of these top 10 salient genes in PrognoScan, a server to search relationship between expression of genes and patients' overall and disease-free survival across an ensemble of

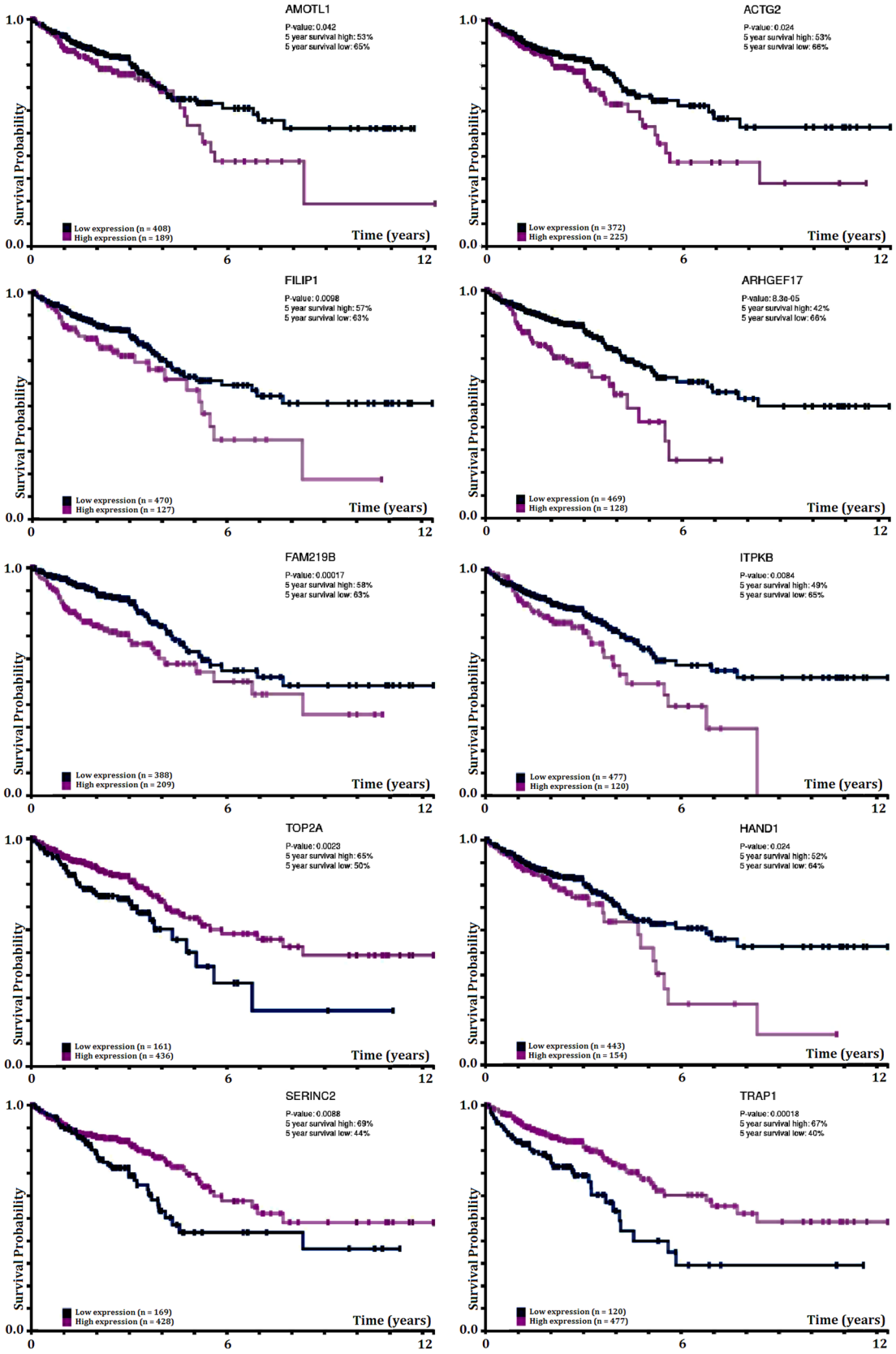


**Figure 7.** Diapeutics implication of top 10 protein-coding salient genes in CRC. Kaplan–Meier (KM) survival analysis of overall survival with respect to expression of top 10 protein-coding salient genes in CRC samples. In each plot, the abscissa represents 'Time in Years' and the ordinate represent 'Survival Probability'. Log-rank p-value for KM plot represents a significant correlation between mRNA expression level and patient survival by exhibiting significant differences in survival between genes' high and low expression. Protein-coding salient genes exhibited statistically significant (log-rank p-value ≤ 0.05) in the overall survival endpoint (refer to Supplementary File Figure S1–S10 for enlarged view).

microarray datasets[43], revealed variation in the results between the microarray datasets (Supplementary spreadsheet). This variation in the results can be attributed (to some extent) to sampling error owing to the small-scale nature of the microarray dataset. Moreover, as most of these genes do not qualify characteristics of a DEG (t-test adjusted p-value of $\leq 0.5$ and Log Fold Change more than 2), they are not considered to be associated with the manifestation of disease outcome when applying conventional microarray analysis technique.

MalaCards[106] is a comprehensive human disease/maladies database integrating data from more than seventy sources, including GeneCards[107] and GeneAnalytics[108]. It provides 'MalaCards InFormaTion (MIFT) Score', which signifies the richness of the gene's information against each disease associated with it; the higher the MIFT score, the more significant the annotation results of the gene is to the disease based on previously published literature[106]. The top 15 genes with the highest LRI score were further evaluated in the MalaCards database[106] to assess their annotation of the disease to the Gene to verify the significance of the results (Supplementary File Table S9).

Though varying MIFT score against major cancer is evident from multiple reports on the involvement in major cancer type, seven (7) genes viz. *MIR143HG*, *AMOTL1*, *FILIP1*, *FAM219B*, *SERINC2*, *APOBR*, *MRPS9* exhibited no MIFT score and no results against CRC (aqua blue rows in Table S9 of Supplementary File). Another seven (7) genes viz. *TRAP1*, *ACTG2*, *HAND1*, *ITPKB*, *PAG1*, *CAMSAP1*, *ARHGEF17* are known to be associated with other cancer types as evident by the existence of prior reports on these genes; yet they exhibited low MIFT score against CRC, suggesting these genes lacks sufficient annotated information on the genes' involvement in CRC owing to the scanty number of the report as strong conclusive evidence to corroborate (olive green rows in Table S9 of Supplementary File). The result implies the novelty in the present work as none-to-scanty reports exist for most top genes in terms of association with CRC. Only *TOP2A* exhibited a high MIFT score against CRC and other cancer types, suggesting strong evidence of prior report on the association with CRC and other cancer types (purple rows in Table S9 of Supplementary File). The *TOP2A* being in the top LRI scoring gene and exhibiting a high MIFT score also suggest that the present work is endorsed by the corroborating work published previously (Supplementary File Table S9). All the genes exhibited a varying degree of results in terms of association with other cancer types.

Previously, graph theory-based work demonstrated using the Human Disease Network and Disease Gene Network that majority of the molecular machinery underlying diseases are highly interconnected– sharing functionally[109,110] as well as their genomic changes[111]. The nature of upstream perturbation in the activity of genes interconnected in a network of complex metabolic pathways can relay diverse perturbation effects in the dynamics of downstream functionality, which may lead to any of the diverse range of (patho) phenotypes associated with downstream pathway's activities[110]. A glance over the MalaCards table reveals that most of these top genes are involved in other cancer subtypes, including breast cancer. A high degree of interconnectedness in genes is observed in CRC with breast cancer and lymphoma. Moreover, both CRC with breast cancer shares many genes with the etiological role[109]. Thus, it is apparent that genes with no or low MIFT score for CRC also exhibited no or low scores for Breast cancer and genes that exhibited high MIFT scores for CRC also demonstrated high scores for Breast Cancer (Table S9 of Supplementary File). Prior reports exhibited that many of these genes are differentially expressed and have a possible etiological role in the manifestation of CRC and Breast Cancer (compared to normal conditions). However, the regulation patterns of these genes in combination (upregulation or downregulation) with respect to CRC and Breast Cancer are not synchronized and lacks coherence[112]. Moreover, the cancer genome 'landscape' suggest varying degree of acquisition of mutation that drives the cell towards CRC or Breast cancer[113].

The LRI method ranks the genes relevance to the condition in an asymmetrical way, with 126 salient genes exhibiting the positive LRI score. The top 10 genes exhibited LRI scores in quartile above the median rank score for these 126 genes (Fig. 1). The extreme values assigned to these genes suggest that their relevance compared to other lower-ranked genes is relatively more, and the relevance of these lower-ranked genes is more than other genes not included in the (Supplementary File Table S1) list. It is well-established that the genes clustering together in expression analysis exhibit common biological ontologies[18,50,114,115]. Hence, predominant functions associated with all the salient genes in a specific cellular context can provide a clear view of affected functional terms in CRC progression. Various GO (Figs. 2 and 3), KEGG (Fig. 4), Reactome (Fig. 5) and PPI (Fig. 6) terms exhibited over-representation than normal in CRC samples characterized by statistically significant low p-value. These salient genes' similar scores can be attributed to the various networks they represent. The significance of the methodology is evident from the observation that relevance of the top 10 protein coding genes as major player with probable etiological role in CRC was also clearly established by the using Kaplan and Meier method of survival analysis (Fig. 7) with significance assessment using Log-Rank tests[50,116]. The novelty of the method can be assessed from the fact that majority (except two) of the salient genes are absent in the existing knowledge database of cancer biomarkers (Fig. 1D). The resulting salient genes showed a dearth of the previous reports in a highly cited and manually curated biomarkers database of repute[30]. The report opens up new dimensions in investigating these salient genes by in vitro and in vivo experiments and ushers new hope in diapeutics by providing novel gene targets for mitigating the development and metastasis of CRC. The report also stresses the algorithm's effectiveness in assessing the importance of individual genes in cancer etiology, utilizing only expression patterns at the molecular scale. This report also presents an opportunity to pounder over the use of non-conventional GT approaches in assessing genes' relevance using genome-wide expression dataset for application in diapeutics to conquer this and other dreaded diseases.

Finally, we can conclude that the mortality caused by cancer can be checked by early diagnostic screening with the help of biomarkers and effective targeted therapeutics. The knowledge discovery of salient genes associated with CRC can fill many voids related to biomarkers, perturbed biochemical pathways, and genes' action during and prior to cancer development. We employed a game-theoretic link relevance Index (LRI) scoring approach on the high-throughput transcriptomics dataset to identify salient genes in CRC. One hundred and twenty-six

(126) salient genes demonstrated a positive LRI score (LRI > 0), indicating the significance of these genes in network games of genes. Investigation of the diverse gene ontology revealed eleven overrepresentations for major Biological processes. GO term for regulation of ion transmembrane transporter activity (GO:0032412) exhibited overrepresentation of the Molecular Function while six overrepresentations were obtained for major Cellular Component. Although considerable enrichment was observed for thirty-eight KEGG pathways and three Reactome pathways, no enrichment was observed for the salient genes concerning chromosome location. The investigation reports the centrality nature of *MiR143HG*, *AMOTL1*, *ACTG2*, *FILIP1*, *ARHGEF17*, *FAM219B*, and other genes in CRC progression, which is lacking in previous studies and public repositories. Furthermore, the resulting information will enhance and supplement the existing knowledge base on CRC and aid future diapeutics investigations. The robustness of the present findings provides the opportunity to re-evaluate the genes associated with diseases and expand the gene-disease databases. The report also highlights LRI algorithms aided genes assessment to evaluate their contribution as a major factor with an etiological role in complex human disease conditions.

## Data availability

The meta-dataset analysed (E-MTAB-6698) during the current study are available in the [Arrayexpress] repository, [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6698/]. All data generated during this study are included in this published article [and its supplementary information files].The in-house script for calculating the LRI is available at https://github.com/Vishwabaruah/GT_LRI.

Received: 1 October 2021; Accepted: 22 July 2022
Published online: 04 August 2022

## References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Guérin, A. *et al.* Risk of developing colorectal cancer and benign colorectal neoplasm in patients with chronic constipation. *Aliment. Pharmacol. Ther.* **40**, 83–92 (2014).
3. Blasi, V. D. *et al.* Major hepatectomy for colorectal liver metastases in patients aged over 80: A propensity score matching analysis. *Dig. Surg.* **35**, 333–341 (2018).
4. Soliman, A. S. *et al.* Colorectal cancer in egyptian patients under 40 years of age. *Int. J. Cancer* **71**, 26–30 (1997).
5. Redmond, J., Vanderpool, R. & McClung, R. Effectively communicating colorectal cancer screening information to primary care providers. *Am. J. Health Educ.* **43**, 194–201 (2012).
6. Lewis, D. R. *et al.* Early estimates of SEER cancer incidence, 2014. *Cancer* **123**, 2524–2534 (2017).
7. Needham, D. *et al.* Bottom up design of nanoparticles for anti-cancer diapeutics: "Put the drug in the cancer's food". *J. Drug Target.* **24**, 836–856 (2016).
8. Cunningham, D. *et al.* Colorectal cancer. *The Lancet* **375**, 1030–1047 (2010).
9. Stein, A., Atanackovic, D. & Bokemeyer, C. Current standards and new trends in the primary treatment of colorectal cancer. *Eur. J. Cancer* **47**, S312–S314 (2011).
10. Bailey, J. R., Aggarwal, A. & Imperiale, T. F. Colorectal cancer screening: Stool DNA and other non-invasive modalities. *Gut Liver* **10**, 204 (2016).
11. Mishra, A. & Verma, M. Cancer biomarkers: Are we ready for the prime time?. *Cancers* **2**, 190–208 (2010).
12. Bhatt, A. N., Mathur, R., Farooque, A., Verma, A. & Dwarakanath, B. S. Cancer biomarkers: Current perspectives. *Indian J. Med. Res.* **132**, 129–149 (2010).
13. Koncina, H. & Rauh, L. Prognostic and predictive molecular biomarkers for colorectal cancer: Updates and challenges. *Cancers* **12**, 319 (2020).
14. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5**, 3231 (2014).
15. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
16. Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* **10**, 25–50 (2015).
17. Seton-Rogers, S. Passengers masquerading as cancer drivers. *Nat. Rev. Cancer* **19**, 485–485 (2019).
18. Sumithra, B., Saxena, U. & Das, A. B. A comprehensive study on genome-wide coexpression network of KHDRBS1/Sam68 reveals its cancer and patient-specific association. *Sci. Rep.* **9**, 11083 (2019).
19. Sun, M. W. *et al.* Game theoretic centrality: A novel approach to prioritize disease candidate genes by combining biological networks with the shapley value. *BMC Bioinform.* **21**, 356 (2020).
20. Moretti, S., Fragnelli, V., Patrone, F. & Bonassi, S. Using coalitional games on biological networks to measure centrality and power of genes. *Bioinformatics* **26**, 2721–2730 (2010).
21. Bora, P. N. *et al.* Identifying the salient genes in microarray data: A novel game theoretic model for the co-expression network. *Diagnostics* **10**, 586 (2020).
22. Lim, S. B., Tan, S. J., Lim, W.-T. & Lim, C. T. Compendiums of cancer transcriptomes for machine learning applications. *Sci. Data* **6**, 194 (2019).
23. Lim, S. B. *et al.* Pan-cancer analysis connects tumor matrisome to immune response. *npj Precis. Oncol.* **3** (2019).
24. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
25. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
26. Chan, B. K. C. Data analysis using R programming. *Adv. Exp. Med. biol.* **1082**, 47–122 (2018).
27. 4.0.0., R. D. C. T. A language and environment for statistical computing. *R Found. Stat. Comput.* **2**, https://www.R-project.org (2020).
28. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
29. Zhao, B. *et al.* Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. *Cancer Control* **26**, 1073274819831260 (2019).
30. Zhang, X. *et al.* Cell marker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2018).
31. Tomczak, A. *et al.* Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* **8**, 1–10 (2018).

32. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705–706 (2016).
33. Shannon, P. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
34. Baruah, V. J. *et al.* Integrated computational approach toward discovery of multi-targeted natural products from thumbai (leucas aspera) for attuning NKT cells. *J. Biomol. Struct. Dyn.* https://doi.org/10.1080/07391102.2020.1844056 (2020).
35. Bindea, G. *et al.* ClueGO: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
36. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).
37. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable proteinprotein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2020).
38. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419–1260419 (2015).
39. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan507 (2017).
40. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
41. Hutter, C. & Zenklusen, J. C. The cancer genome atlas: Creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
42. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
43. Mizuno, H., Kitada, K., Nakai, K. & Sarai, A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med. Genomics* **2**, 18 (2009).
44. Parkinson, D. R. *et al.* Evidence of clinical utility: An unmet need in molecular diagnostics for patients with cancer. *Clin. Cancer Res.* **20**, 1428–1444 (2014).
45. Sawyers, C. L. & Veer, L. J. Reliable and effective diagnostics are keys to accelerating personalized cancer medicine and transforming cancer care: A policy statement from the American association for cancer research. *Clin. Cancer Res.* **20**, 4978–4981 (2014).
46. Kuipers, E. J. *et al.* Colorectal cancer. *Nat. Rev. Dis. Prim.* **1**, 15065 (2015).
47. Poste, G. Bring on the biomarkers. *Nature* **469**, 156–157 (2011).
48. Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. Cancer biomarker discovery and validation. *Transl. Cancer Res.* **4**, 256–269 (2015).
49. Hammond, W. A., Swaika, A. & Mody, K. Pharmacologic resistance in colorectal cancer: A review. *Ther. Adv. Med. Oncol.* **8**, 57–84 (2015).
50. Jayanthi, V. S. P. K. S. A., Das, A. B. & Saxena, U. Grade-specific diagnostic and prognostic biomarkers in breast cancer. *Genomics* **112**, 388–396 (2020).
51. Moretti, S. & Vasilakos, A. V. An overview of recent applications of game theory to bioinformatics. *Inf. Sci.* **180**, 4312–4322 (2010).
52. Lucchetti, R., Moretti, S., Patrone, F. & Radrizzani, P. The Shapley and Banzhaf values in microarray games. *Comput. Oper. Res.* **37**, 1406–1412 (2010).
53. Fragnelli, V. & Moretti, S. A game theoretical approach to the classification problem in gene expression data analysis. *Comput. Math. Appl.* **55**, 950–959 (2008).
54. Moretti, S., Patrone, F. & Bonassi, S. The class of microarray games and the relevance index for genes. *TOP* **15**, 256–280 (2007).
55. Moretti, S. Game theory applied to gene expression analysis. *4OR* **7**, 195–198 (2008).
56. Cesari, G., Algaba, E., Moretti, S. & Nepomuceno, J. A. An application of the shapley value to the analysis of co-expression networks. *Appl. Netw. Sci.* **3**, 35 (2018).
57. Albino, D. *et al.* Identification of low intratumoral gene expression heterogeneity in neuroblastic tumors by genome-Wide expression analysis and Game Theory. *Cancer* **113**, 1412–1422 (2008).
58. Esteban, F. J. & Wall, D. P. Using game theory to detect genes involved in Autism Spectrum Disorder. *TOP* **19**, 121–129 (2011).
59. Cesari, G., Algaba, E., Moretti, S. & Nepomuceno, J. A. A game theoretic neighbourhood-based relevance index. *Stud. Comput. Intell.* **689**, 29–40 (2018).
60. Hernández-Ochoa, E. O., Pratt, S. J. P., Lovering, R. M. & Schneider, M. F. Critical role of intracellular RyR1 calcium release channels in skeletal muscle function and disease. *Front. Physiol.* **6**, 420 (2015).
61. Poujade, M. *et al.* Collective migration of an epithelial monolayer in response to a model wound. *Proc. Natl. Acad. Sci. USA* **104**, 15988–15993 (2007).
62. Vitorino, P., Hammer, M., Kim, J. & Meyer, T. A steering model of endothelial sheet migration recapitulates monolayer integrity and directed collective migration. *Mol. Cell. Biol.* **31**, 342–350 (2011).
63. Christensen, J. F. *et al.* Muscle dysfunction in cancer patients. *Ann. Oncol.* **25**, 947–958 (2014).
64. Coletti, D. Chemotherapy-induced muscle wasting: an update. *Eur. J. Transl. Myol.* **28**, 7587 (2018).
65. Al-Majid, S. & Waters, H. The biological mechanisms of cancer-related skeletal muscle wasting: The role of progressive resistance exercise. *Biol. Res. Nurs.* **10**, 7–20 (2008).
66. van Waart, H. *et al.* Effect of low-intensity physical activity and moderate- to high-intensity physical exercise during adjuvant chemotherapy on physical fitness, fatigue, and chemotherapy completion rates: Results of the PACES randomized clinical trial. *J. Clin. Oncol.* **33**, 1918–1927 (2015).
67. Barreto, R. *et al.* Chemotherapy-related cachexia is associated with mitochondrial depletion and the activation of ERK1/2 and p38 MAPKs. *Oncotarget* **7**, 43442–43460 (2016).
68. Barreto, R. *et al.* Cancer and chemotherapy contribute to muscle loss by activating common signaling pathways. *Front. Physiol.* **7**, 472 (2016).
69. Ordóñez, N. G. Podoplanin: a novel diagnostic immunohistochemical marker. *Adv. Anat. Pathol.* **13**, 83–88 (2006).
70. Hannon, E. *et al.* A role for CaV1 and calcineurin signaling in depolarization-induced changes in neuronal DNA methylation. *Neuroepigenetics* **3**, 1–6 (2015).
71. Song, L. *et al.* Calsequestrin 2 (CASQ2) mutations increase expression of calreticulin and ryanodine receptors, causing catecholaminergic polymorphic ventricular tachycardia. *J. Clin. Investig.* **117**, 1814–1823 (2007).
72. Gélébart, P. *et al.* Expression of endomembrane calcium pumps in colon and gastric cancer cells. Induction of SERCA3 expression during differentiation. *J. Biol. Chem.* **277**, 26310–26320 (2002).
73. Naydenov, N. G., Lechuga, S., Huang, E. H. & Ivanov, A. I. Myosin motors: Novel regulators and therapeutic targets in colorectal cancer. *Cancers (Basel)* **13**, 1–24 (2021).
74. Yarla, N., Madka, V. & Rao, C. Targeting triglyceride metabolism for colorectal cancer prevention and therapy. *Curr. Drug Targets* https://doi.org/10.2174/1389450122666210824150012 (2021).
75. Brown, R. E., Short, S. P. & Williams, C. S. Colorectal cancer and metabolism. *Curr. Colorectal Cancer Rep.* **14**, 226–241 (2018).
76. Phillips, B. E. *et al.* Effect of colon cancer and surgical resection on skeletal muscle mitochondrial enzyme activity in colon cancer patients: A pilot study. *J. Cachexia. Sarcopenia Muscle* **4**, 71–77 (2013).
77. Anderson, K. J., Cormier, R. T. & Scott, P. M. Role of ion channels in gastrointestinal cancer. *World J. Gastroenterol.* **25**, 5732–5772 (2019).

78. Fnu, G. & Weber, G. F. Alterations of ion homeostasis in cancer metastasis: Implications for treatment. *Front. Oncol.* **11**, 765329 (2021).
79. Vacante, F., Denby, L., Sluimer, J. C. & Baker, A. H. The function of miR-143, miR-145 and the MiR-143 host gene in cardiovascular development and disease. *Vascul. Pharmacol.* **112**, 24–30 (2019).
80. Xie, H. *et al.* LncRNA miR143HG suppresses bladder cancer development through inactivating wnt/β-catenin pathway by modulating miR-1275/AXIN2 axis. *J. Cell. Physiol.* **234**, 11156–11164 (2018).
81. Lin, X. *et al.* Long non-coding RNA miR143HG predicts good prognosis and inhibits tumor multiplication and metastasis by suppressing mitogen-activated protein kinase and WNT signaling pathways in hepatocellular carcinoma. *Hepatol. Res.* **49**, 902–918 (2019).
82. Lv, M. *et al.* Angiomotin family members: Oncogenes or tumor suppressors?. *Int. J. Biol. Sci.* **13**, 772–781 (2017).
83. Huang, T. *et al.* The physiological role of motin family and its dysregulation in tumorigenesis. *J. Transl. Med.* **16** (2018).
84. Couderc, C. *et al.* AMOTL1 promotes breast cancer progression and is antagonized by merlin. *Neoplasia (United States)* **18**, 10–24 (2016).
85. Ou, R. *et al.* circAMOTL1 motivates AMOTL1 expression to facilitate cervical cancer growth. *Mol. Therapy Nucleic Acids* **19**, 50–60 (2020).
86. Edfeldt, K., Hellman, P., Westin, G. & Stalberg, P. A plausible role for actin gamma smooth muscle 2 (ACTG2) in small intestinal neuroendocrine tumorigenesis. *BMC Endocr. Disord.* **16**, 19 (2016).
87. James, K. N. *et al.* Expanding the genotypic spectrum of ACTG2-related visceral myopathy. *Mol. Case Stud.* **7**, a006085 (2021).
88. Wu, Y. *et al.* Identification of ACTG2 functions as a promoter gene in hepatocellular carcinoma cells migration and tumor metastasis. *Biochem. Biophys. Res. Commun.* **491**, 537–544 (2017).
89. Adammek, M. *et al.* MicroRNA miR-145 inhibits proliferation, invasiveness, and stem cell phenotype of an in vitro endometriosis model by targeting multiple cytoskeletal elements and pluripotency factors. *Fertil. Steril.* **99**, 1346-1355.e5 (2013).
90. Beck, A. H. *et al.* Discovery of molecular subtypes in leiomyosarcoma through integrative molecular profiling. *Oncogene* **29**, 845–854 (2009).
91. Lauvrak, S. U. *et al.* Functional characterisation of osteosarcoma cell lines and identification of mRNAs and miRNAs associated with aggressive cancer phenotypes. *Br. J. Cancer* **109**, 2228–2236 (2013).
92. Drew, J. E. *et al.* Predictive gene signatures: Molecular markers distinguishing colon adenomatous polyp and carcinoma. *PLoS ONE* **9**, e113071 (2014).
93. Edfeldt, K. *et al.* Different gene expression profiles in metastasizing midgut carcinoid tumors. *Endocr. Relat. Cancer* **18**, 479–489 (2011).
94. Kwon, M. *et al.* Functional characterization of filamin a interacting protein 1like, a novel candidate for antivascular cancer therapy. *Can. Res.* **68**, 7332–7341 (2008).
95. Kwon, M. *et al.* Reduced expression of FILIP1L, a novel WNT pathway inhibitor, is associated with poor survival, progression and chemoresistance in ovarian cancer. *Oncotarget* **7**, 77052–77070 (2016).
96. Isokane, M. *et al.* ARHGEF17 is an essential spindle assembly checkpoint factor that targets Mps1 to kinetochores. *J. Cell Biol.* **212**, 647–659 (2016).
97. García, I.,. Lysophosphatidic acid promotes lung carcinoma cell migration via ARHGEF17, a RhoGEF directly controlled by g. *FASEB J.* **34**, 1 (2020).
98. Jikuya, H. *et al.* Characterization of long cDNA clones from human adult spleen. II. The complete sequences of 81 cDNA clones. *DNA Res.* **10**, 49–57 (2003).
99. Kimura, K. *et al.* Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).
100. Heestand, G. M., Schwaederle, M., Gatalica, Z., Arguello, D. & Kurzrock, R. Topoisomerase expression and amplification in solid tumours: Analysis of 24,262 patients. *Eur. J. Cancer* **83**, 80–87 (2017).
101. Coss, A. *et al.* Increased topoisomerase II α expression in colorectal cancer is associated with advanced disease and chemotherapeutic resistance via inhibition of apoptosis. *Cancer Lett.* **276**, 228–238 (2009).
102. Tan, J. *et al.* Integrative epigenome analysis identifies a polycomb-targeted differentiation program as a tumor-suppressor event epigenetically inactivated in colorectal cancer. *Cell Death Dis.* **5**, e1324–e1324 (2014).
103. da Cunha, J. P. C. *et al.* The human cell surfaceome of breast tumors. *Biomed. Res. Int.* **2013**, 1–11 (2013).
104. Chen, J., Wang, Z., Shen, X., Cui, X. & Guo, Y. Identification of novel biomarkers and small molecule drugs in human colorectal cancer by microarray and bioinformatics analysis. *Mol. Genet. Genom. Med.* **7**, e00713 (2019).
105. Ghaffari, S. *et al.* An integrated multi-omics approach to identify regulatory mechanisms in cancer metastatic processes. *Genome Biol.* **22**, 19 (2021).
106. Rappaport, N. *et al.* MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* **45**, D877–D887 (2017).
107. Safran, M. *et al.* The GeneCards Suite. in *Pract. Guid. To life sci. databases* 27–56 (Springer Singapore, 2021). https://doi.org/10.1007/978-981-16-5812-9_2.
108. Fuchs, S. B. A. *et al.* GeneAnalytics: An integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. *OMICS J. Integr. Biol.* **20**, 139–151 (2016).
109. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
110. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
111. Barrenäs, F. *et al.* Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol.* **13**, R46 (2012).
112. Liu, T., Zhou, L., Li, D., Andl, T. & Zhang, Y. Cancer-associated fibroblasts build and secure the tumor microenvironment. *Front. cell Dev. Biol.* **7**, 60 (2019).
113. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
114. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863–14868 (1998).
115. Reynier, F. *et al.* Importance of correlation between gene expression levels: Application to the type i interferon signature in rheumatoid arthritis. *PLoS ONE* **6**, e24828 (2011).
116. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).


## Acknowledgements

VJB and BS thankfully acknowledge research grants provided by the Department of Biotechnology, Government of India (Grant No. BT/PR25099/NER/95/1014/2017). SB and RK are greatly indebted to UK-India Education and Research Initiative (UKIERI) for their generosity and financial support (Grant No. 184-15/2017(IC)). SB also extends appreciation and acknowledgement for the financial support received from Assam Science Technology & Environment Council (ASTEC), Government of Assam (Grant No. ASTEC/S&T/192(171)/2019-20/2762).

## Author contributions

V.J.B. and P.N.B. conceived and designed the experiments. V.J.B., P.N.B., P.M., A.S. and B.S. performed the experiments. V.J.B., P.N.B. and B.S. analyzed the data. V.J.B. and B.S. rendered the figures. V.J.B., B.S., and S.B. wrote the manuscript. All authors critically revised the intellectual content of the manuscript.

## Funding

This research was funded by UK-India Education and Research Initiative (UKIERI) (Grant Number 184-15/2017(IC)) and Assam Science and Technology and Environment Council (Grant No. ASTEC/S&T/192(171)/2019-20/2762). A part of the research was carried out using the assets created by the grant received from the Department of Biotechnology, Govt. of India (Grant No. BT/PR25099/NER/95/1014/2017).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-17266-0.

**Correspondence** and requests for materials should be addressed to V.J.B. or S.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.